Final Report
HumRRO
FR-01-26
June, 2001

The Accuracy of School Classifications for the Interim Accountability Cycle of the Kentucky Commonwealth Accountability and Testing System

R. Gene Hoffman
Lauress L. Wise

Prepared for:
Kentucky Department of Education
Frankfort, KY 40601

Contract Number M-00003669

Human Resources Research Organization
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

THE ACCURACY OF SCHOOL CLASSIFICATIONS FOR THE INTERIM ACCOUNTABILITY CYCLE OF THE KENTUCKY COMMONWEALTH ACCOUNTABILITY AND TESTING SYSTEM

R. GENE HOFFMAN AND LAURESS L. WISE
HUMAN RESOURCES RESEARCH ORGANIZATION

Introduction

Kentucky's Commonwealth Accountability and Testing System (CATS) was implemented in 1999 as a modification of the Kentucky Instructional Results Information System (KIRIS). Beginning with KIRIS, public schools in Kentucky have been classified by their success in educating their students. Both the KIRIS and CATS systems have tied significant consequences to schools' classifications. This report presents an analysis of the accuracy of the school classifications for the Interim Accountability cycle that bridged the KIRIS program to the CATS program.

It is well known that achievement test scores cannot be perfect. Consequently, we can infer that CATS school classifications, which are derived primarily from students' achievement test scores, will not be perfect either. It is important, therefore, to document the accuracy, not only of student achievement scores, but also of school-level averages for test scores, and the resulting accuracy in school classifications. Hoffman and Wise (2000a, 2000b) have previously documented student-level accuracy. This report focuses on school-level accuracy.

The most technical aspects of our analyses are reserved for the Technical Appendix. In the following sections, we have attempted to make the conclusions accessible to an audience with a "working" knowledge of testing. The information is, however, technical by nature. Further simplification may lead to misrepresentation of the results.

Brief Background on KIRIS and CATS[1]

In order to understand the analyses presented in this report, it is necessary to understand the accountability computations that generate school classifications for the CATS Interim Accountability Cycle. The Interim Cycle is a hybrid created to bridge between an old system and a new system. As a result, there are some aspects of the analyses of the Interim Cycle that apply neither to the old KIRIS methods for classifying schools nor to the future CATS method.

Beginning with KIRIS, students in selected grades took tailored tests in seven different subject areas and completed writing portfolios. Under KIRIS, schools were then classified according to increases in average student achievement over a four-year period. Specifically, average scores for student cohorts during the first two years of a four-year cycle were used to

[1] This brief overview is intended for those unfamiliar with KIRIS and CATS.

calculate target scores for the student cohorts who would be passing through the school during the third and fourth years of the cycle. Under KIRIS, schools were rewarded or sanctioned depending on whether they met their targets or not. CATS will continue a similar school-level accountability model using the Kentucky Core Content Test, a revision of the old KIRIS test. Every two years, achievement scores of students within the school will be compared to target scores based on the achievement scores of past students. For the duration of CATS, schools will be expected to continually increase the average scores of the student cohorts passing through their doors.

CATS includes eight assessments administered to selected grades such that all eight assessments are administered in a typical elementary school, a typical middle school, and a typical high school. Table 1 indicates the grades in which the assessments are administered. With each assessment, students are classified into one of four achievement levels: Novice, Apprentice, Proficient, and Distinguished. For the four primary content disciplines (Reading, Mathematics, Science, and Social Studies), the lower two levels, Novice and Apprentice, are subdivided into thirds (low, middle, and high), resulting in eight achievement categories. Based on Kentucky statutes, points are awarded to these eight categories in the following array (from low Novice to Distinguished): 0, 13, 26, 40, 60, 80, 100, and 140. For the remaining content areas, including Arts & Humanities, Practical Living/Vocational Studies, Writing (including the on-demand writing prompt and writing portfolios), scores are limited to two levels of Novice (with 0 points for students who make no attempt to answer and 13 points for those who try) and one level each for Apprentice, Proficient, and Distinguished (at 60, 100, and 140 points, respectively). The point values are used to calculate schools' average student achievement in each content area.

Table 1
Weighted Assessments in the Academic and Non-Academic Components of the CATS School Accountability Component 1 Index

Academic Content	School Level and Grade						
	Elementary		Middle		High		
	4	5	7	8	10	11	12
Reading	.20		.15		.15		
Mathematics		.20		.15		.15	
Science	.15		.15			.15	
Social Studies		.15		.15		.15	
Arts & Humanities		.05		.075		.075	
Practical Living/ Vocational Studies		.05		.075	.075		
Writing Prompt	.03		.03				.03
Writing Portfolio	.12		.12				.12
Non-Academic	.05		.10		.10		

Schools also receive scores for a composite of non-academic factors such as attendance rate, retention rate, and dropout rate. The academic and non-academic scores are combined to form Component 1 of the CATS accountability index. Table 1 also indicates the weights used to combine school average scores on each assessment into the Component 1 total score.

Salient Features of CATS for the Interim Accountability Cycle

The transition from KIRIS to CATS ushered in a number of changes, most importantly the inclusion of multiple-choice items and changes in the computation of school scores from student scores. As a result, school indexes in 1999 and 2000 could not be directly compared to scores from previous years. The Interim Cycle is therefore a one-cycle deviation from the straightforward school improvement concept that was used in KIRIS and will be used by CATS in future accountability cycles.

The CATS Interim Accountability Cycle began with the school year of 1996-1997 and ended with the school year 1999-2000. Because testing occurred in the spring of each school year, this report will reference each year with the spring date only. Data for 1997 and 1998 constitute the "base years" upon which target scores were set for the "final years" of 1999 and 2000. For the interim cycle only, "Improvement Goals" were set using a regression approach in which the average of 1997 and 1998 Component 1 indexes were used to make a statistical prediction for the average of 1999 and 2000 Component 1 indexes. Schools that performed better than their predicted performance were rewarded. Schools that performed below their predicted scores by more than one standard error of prediction were offered assistance. The one standard error of prediction provided a safety band to reduce instances of schools being given the stigma of substandard performance by chance alone. Schools with indexes between their predicted scores and one standard deviation below their predicted scores were designated "Maintaining" with no further consequences. Details of the regression model computations are available elsewhere (Carlson, 1999).

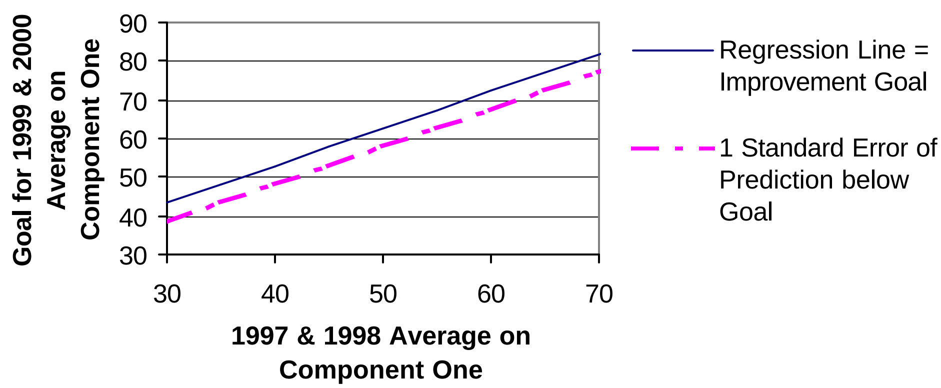

Figure 1. Interim Accountability Model: Schools with actual 1999/2000 scores above the solid line were rewarded. Schools below the dashed line were offered assistance. School in between were labeled "Maintaining."

Figure 1 depicts the regression line that provides the improvement goal for 1999/2000 based on 1997 and 1999 student achievement. The dotted line indicates one standard error below target performance set by the regression line. Figure 1 is based on data for elementary schools; however, the model is the same for middle and high schools with the placement of the regression line being altered slightly based on separate regression analyses for the three different levels of school. The handicapping concept is evident. Compared to initially high-scoring schools, initially low-scoring schools were not required to have as high final years' scores in order to receive rewards. Conversely, in order to avoid being labeled as needing assistance, initially high-scoring schools must have attained higher final years' scores than initially low-scoring schools.

The classification decision process makes reference to the "standard error of prediction." Error of prediction is the extent to which schools could vary from their regression-based prediction by chance alone, *all other factors being equal*. From the perspective of the accountability system, differences beyond some degree of chance signal that "all other factors" were not equal. Based on this probability concept, schools that were one standard error below their predicted level for 1999 and 2000 were judged to be performing differently from the trend of all schools, and were, therefore, viewed as likely to be in need of instructional assistance.

For the regression model to work, the relationship between past and future performance can be neither too low nor too high. In order for traditionally low scoring schools with challenging populations to have had a chance at receiving rewards, the regression model depended on there being a reasonably strong correlation between past and future performance (i.e., an upward slope in the prediction line in Figure 1). Such a relationship served as a handicapping system in which each school's performance was compared to other schools that had similar scores during the base years. On the other hand, a very strong relationship between past and future performance across all schools would have meant that there was little if any true difference among schools in their instructional improvements, and therefore, no real basis for determining school classifications. Fortunately, Interim Cycle correlations at the elementary, middle, and high school levels appeared to be about right. The observed correlations, which will be presented below, were strong enough to provide a handicapping system, but not so strong as to automatically conclude that there were no true differences in instructional improvements among schools.

There is an important caveat to this interpretation, however. All correlations are affected by measurement accuracy. Measurement error makes correlations an underestimate of true relationships. Therefore, it is possible for the Interim Cycle to have the appearance of effectively sorting strong and weak schools when it did not actually do so. That is, the true, underlying relationship between performance in the base and final years may have been so strong that there were no true differences among the schools. Measurement error could have been large enough to reduce the observable relationship, creating an illusion of differences among schools. If this was the case, classifications would have been made on measurement error and, therefore, would have been essentially random.

The purpose of this report is to track the effects of measurement error on the Interim Accountability decisions. Specifically, the objective for the school classification analysis was to estimate reliability and standard errors of *measurement* for schools' interim cycle accountability indexes. Standard errors of measurement were applied to each school's observed index and used to calculate the probability that a school's true performance could have been in a different school classification.

Note the shift in terminology. Standard errors of *measurement* are indicators of error due solely to testing inaccuracy. Standard errors of *prediction* are indicators of the strength of the observed relationship between past performance and future performance. For our purposes, two factors affect the strength of that observed relationship. One is the possibility that schools do actually differ in their instructional effectiveness, such that two different schools with equal starting places will end up with different levels of student achievement. To the extent that there are true differences among schools, the strength of the true relationship between past and future performance will be lower, and consequently, the strength of the observed relationship will be lower. The other factor affecting standard error of prediction is measurement error. Measurement error affects the observation process, reducing our ability to observe the true relationship. The distinction is important: This report focuses on the measurement process and the extent to which it supported classification accuracy for the Interim Accountability Cycle.

A Word about the Analysis Process

The methodology for this classification accuracy analysis was developed by Hoffman and Wise (2000d) and presented to Kentucky's National Technical Advisory Panel on Assessment and Accountability (NTAPAA) on June 22, 2000. Preliminary results were presented during the January meeting with one suggested revision.[2] This report conforms to the NTAPAA approved specifications. The method merges a generalizability procedure patterned after Yen (1997) and Miller (1999) with a classification accuracy procedure developed for the Kentucky Department of Education (Hoffman and Wise, 1999), reviewed by the NTAPAA on two occasions (September 9-10, 1999 and December 16-17, 1999), and presented at the National Council of Measurement in Education annual meeting (Hoffman and Wise, 2000c). Computation details are reserved for the Technical Appendix to this report. However, there are a few key points necessary to understand the results.

First, we conducted separate analyses for the standard configurations of elementary, middle, and high schools such that each configuration included the tested grades (see Table 1).

Second, measurement error is affected by the amount of data available for a particular school: The more data, the less error. As a result of this principle, we expected large schools to be measured more accurately than small schools because their index scores were based on more students. Therefore, we conducted our analyses on three representative sizes of school, selecting schools to represent the lower third in size, the middle third, and the upper third. Representative sizes were selected independently for elementary, middle, and high schools.

Third, similar to the calculation of school accountability indexes, school classification accuracy calculations began with student level data. Using the point values designated for students' achievement level classifications, standard errors of measurement and reliability calculations were first made on each of the eight assessments for each of the three representative sizes for each of the three levels of schools (elementary, middle, and high). These 72 calculations (8 x 3 x 3) were conducted by a procedure known as generalizability theory. An explanation of this procedure and the complete results for each of the 72 calculations are presented in the Technical Appendix. In brief, the generalizability procedure considered the effects of years, students, and test structure in generating results. The generalizability results were used to

[2] The suggestion was to explore the impact of various assumptions about the reliability of the Non-Academic Index. This is described later in the report.

calculate standard error of measurements, reliability, and classification accuracy for each level and representative school size.

Fourth, because of changes from KIRIS to CATS, we were not able to directly estimate similar standard error and reliability estimates for the base years' school index. Instead, we synthesized estimates for the base years from the results of the final years using what we knew about differences in the structure of the tests. (Details are in the Technical Appendix).

Finally, we had no method for estimating the reliability of the Non-Academic scores. We explored using the values 1 (perfect reliability) and 0 (total unreliability) and found that the estimate of overall Component 1 school error was only slightly different from each other. We therefore selected a conservative reliability estimate (.7) to use in the following calculations.

Reliability and Classification Accuracy

The following results actually focus on two indexes. One index is the school-level Component 1 scores averaged for 1999 and 2000, which we refer to as the "final index." The other index is the difference between the final index and the improvement goal, described earlier. Technical speaking, because this difference is derived from a regression equation, the difference is called a "residual." In this technical language, schools with positive residuals (final scores larger than their improvement goals) were rewarded. Schools with negative residuals had final Component 1 scores below their predicted performance. Because of the nature of measurement error, difference scores, such as the residual, tend to have noticeably lower reliability than either of the scores used to compute the difference. Ironically, when the two input scores are highly correlated (such as required by the handicapping objective for using the regression model), then the reliability of the resulting difference can suffer a large decrement. As a result, measurement error in the index for the final years or in the index for the base years would be greatly magnified in the residual. Therefore, we present results for both the final year and the residual in order to gain an understanding of the measurement error in a single score and how computing differences between scores compounds that measurement error.

Table 2 summarizes the results of our analyses. The top portion of the table presents results for the 1999/2000 final-years' Component 1 index. [3] The bottom portion presents results for the residual index, which, again, is the difference between the 1999/2000 final years' index and the improvement goal computed from the 1997/1998 base-years' index.

Columns (a) and (b) identify school level and school size. Size is the number of students in a grade level. Column (c) presents reliability estimates for the 1999/2000 Component 1 index. These reliabilities are all high, with the lowest at .965 and the rest at or above .980. As expected, the index scores for larger schools tend to have higher reliabilities. Note that we report the reliabilities to the third decimal not because we believe that they are that accurate, but to avoid rounding the highest reliability to 1.00, an improbable figure. On the other hand, as high as these reliabilities appear, they are not particularly surprising given the large amount of student data that contributes to the school scores.

[3] When referencing an accountability index that combined two years of data, we will identify the index as either a "1997/1998" index or a "1999/2000" index, indicating the two years that the index includes.

Table 2
Interim Model Reliability Analysis

(a) Grade	(b) Size	1999/2000 Component 1 Index (c) Reliability	(d) Standard Error of Measurement	(e) Classification Accuracy Assuming No Error in Improvement Goal	(f) Correlation between 97/98 and 99/00
Elementary	Small-24	0.965	1.47	83.7%	
Elementary	Medium-60	0.988	0.96	91.0%	0.80
Elementary	Large-96	0.990	0.82	93.1%	
Middle	Small-36	0.988	1.07	84.7%	
Middle	Medium-120	0.988	0.65	94.4%	0.915
Middle	Large-240	0.996	0.47	92.9%	
High	Small-60	0.980	0.94	85.2%	
High	Medium-168	0.993	0.61	88.6%	0.905
High	Large-240	0.995	0.52	88.9%	
Grade	**Size**	**Residual = Accountablity Index (g) Reliability**	**(h) Standard Error of Measurement**	**(i) Classification Accuracy Assuming Measurement Error in Improvement Goal**	**(j) Assistance Point (-1 Standard Error of Regression)**
Elementary	Small-24	0.786	2.22	75.7%	
Elementary	Medium-60	0.927	1.30	87.7%	-4.8
Elementary	Large-96	0.941	1.17	89.6%	
Middle	Small-36	0.811	1.22	82.6%	
Middle	Medium-120	0.800	1.25	83.7%	-3.2
Middle	Large-240	0.933	0.73	88.1%	
High	Small-60	0.727	1.64	73.7%	
High	Medium-168	0.899	1.00	81.9%	-3.2
High	Large-240	0.934	0.81	84.9%	

Standard errors of measurement for the 1999/2000 Component 1 index, Column (d), present a gauge for the amount of measurement error on the same scale as the index. As noted above, a school's test score will vary by chance error. If it were possible to repeatedly measure a school, the new scores would be no more than one standard error higher or lower than the observed score about 67% of the time. In other words, we can be about 67% confident that schools' true scores are within + or – one standard error of their observed scores. Small elementary and small middle schools were the least accurately assessed, with a 67% confidence that they were accurate to within about 1.5 and 1.1 points. For the remaining schools, the standard errors of measurement indicate that schools' scores had a 67% probability of being accurate to within at least 1 point.

Column (e) presents classification accuracy estimates given an assumption that the only measurement error in the accountability system was in the 1999 and 2000 assessments. These percentages represent the probability that a school's true performance was actually in the same category as assigned by the accountability system. Conversely, 100 minus the listed percentages indicate the probability of a school being incorrectly classified based on measurement error in the 1999/2000 Component 1 index, assuming no measurement error in the base years. We will have more to say about interpreting errors later.

Column (f) indicates the correlations between the 1997/1998 Component 1 index and the 1999/2000 Component 1 index. These correlations are strong enough to signal that a handicapping system was operating in the Interim Accountability model. They are also large enough to generate concern about the extent to which measurement affected the classifications.

Columns (g), (h), and (i) present parallel reliabilities, standard errors of measurement, and classification accuracy for the difference between predicted and actual performance, i.e., the accountability residual. These results make the reasonable assumption that measurement error did exist in the base years as well as in the final years. Because of the nature of difference scores, the reliabilities are lower and the standard errors of measurement are higher than those in the top of the table. Standard errors range from just under 1 to just over 2. On the other hand, the standard errors of prediction, given in column (j) are generally about twice as large as the standard errors of measurement. Combined, columns (h) and (j) indicate that some of the variation of schools' observed 1999/2000 indexes from their target indexes was due to measurement error and some of it was potentially related to factors such as true differences in instructional effectiveness.

Column (i) indicates the accuracy of the school classifications when measurement error in both the base years and the final years is considered. These figures indicate accuracy rates between approximately 75% and 90%. The nature of classification is such that errors are inevitable. The nature of these errors will be explored in the next two sections of the report.

Classification accuracy by assigned classification

Tables 3 through 11 provide more detailed classification statistics and indicate where the classification errors are likely to occur.

The percentages in Tables 3 through 11 represent the accuracy for each of the nine level and size categories for schools. The italicized numbers are expected percentages of schools within the given category. In each table, the sum of all of the italicized percents is 100. The "Total Assigned" row indicates the percent of schools that were assigned each of the three accountability classifications. In Table 3, for example, 21.6% of the schools representing the small elementary category had assessment scores that placed them in the "Needs Assistance" category. Likewise, 27.0% were placed in the "Maintaining Category" and the remaining 51.4% were placed in the "Meets Goal" category.

Table 3.
Small Elementary Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***15.0%***	*3.5%*	*0.1%*	18.7%
Maintaining	*6.5%*	***17.9%***	*8.4%*	32.8%
Meets Goal	*0.1%*	*5.7%*	***42.8%***	48.5%
Total Assigned:	21.6%	27.0%	51.4%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 75.7%

Test scores are not perfect indicators of true student achievement. Therefore, some proportion of schools are expected to have true student achievement that places the schools in categories that match their assigned categories. Another proportion of schools are expected to have true student achievement that places the schools in categories other than their assigned categories. The bold numbers in Table 3 indicate the expected percentages of accurate classifications. That is, approximately 15% of all small elementary schools are expected to be accurately classified as "Needs Assistance," approximately 18% of all schools are expected to be accurately classified as "Maintaining," and approximately 43% of all schools are expected to be accurately classified as "Meets Goal." The sum of these three percentages, 75.7%, gives the approximate percent of all schools expected to be accurately classified given measurement error. That is, based on their less-than-perfect test scores, about 76% of all small elementary schools would be expected to be assigned to the same category of proficiency as would be expected if we actually knew their true achievement.

The non-bold, italicized numbers in Table 3 indicate the proportions of schools that, because of measurement error, are expected to have true achievement classifications different than the classification assigned by the regression formula. For example, about 6% of all small representative elementary schools are expected to have obtained accountability indexes that place them in the "Needs Assistance" range when their true achievement would place them one category higher in the "Maintaining" category. In other words, 6% of small elementary schools may have inadvertently been offered assistance due to measurement error in the system. Conversely, about 3% of all schools are expected to have obtained index scores that place them in the "Maintaining" category, while their true achievement would place in the "Needs Assistance" category. That is, approximately 3% of the small elementary schools should have been offered assistance but were not. Note that the chances were very small (.1%) that a school assigned as "Needing Assistance" was measured so inaccurately that it failed to receive a reward had its true performance been known. Similarly, the chances were very small (.1%) that any school was given rewards for meeting their goal when it would have been offered assistance had its true performance actually been known.

Table 3 also shows that about 6% of the small elementary schools may have deserved rewards for meeting their goal, but because of measurement error, they were assigned to the

"Maintaining" category. Similarly, an estimated 8% of the small elementary schools received rewards when their true performance would have indicated "Maintaining."

The final column indicates the percentages of small elementary schools that might have been assigned to each accountability classification if their students' performance had been measured without error.

Reliability, standard error of measurement, and classification accuracy results for the remaining level and size combinations are presented in Tables 4 through 11. The results may be interpreted in the same manner as above. Table 3 represents the worst case because it has the smallest class sizes. Accuracy for the remaining categories of schools is higher.

Table 4.
Medium Size Elementary Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***6.5%***	*3.7%*	*0.0%*	10.2%
Maintaining	*1.6%*	***27.9%***	*4.5%*	34.0%
Meets Goal	*0.0%*	*2.6%*	***53.2%***	55.8%
Total Assigned:	8.1%	34.2%	57.7%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 87.7%

Table 5.
Large Elementary Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***5.2%***	*1.5%*	*0.0%*	6.7%
Maintaining	*1.4%*	***21.4%***	*4.2%*	27.0%
Meets Goal	*0.0%*	*3.3%*	***63.0%***	66.3%
Total Assigned:	6.6%	26.2%	67.2%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 89.6%

Table 6. Small Middle Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications				
Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***9.1%***	*6.4%*	*0.0%*	15.6%
Maintaining	*0.9%*	***23.2%***	*7.2%*	31.2%
Meets Goal	*0.0%*	*2.9%*	***50.3%***	53.2%
Total Assigned:	10.0%	32.5%	57.5%	100.0%
Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums. Total congruence (sum of bold numbers) = 82.6%				

Table 7. Medium Size Middle Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications				
Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***8.6%***	*4.0%*	*0.0%*	12.6%
Maintaining	*0.8%*	***34.3%***	*3.0%*	38.0%
Meets Goal	*0.0%*	*8.6%*	***40.8%***	49.4%
Total Assigned:	9.4%	46.9%	43.8%	100.0%
Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums. Total congruence (sum of bold numbers) = 83.7%				

Table 8. Large Middle Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications				
Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***3.7%***	*1.9%*	*0.0%*	5.7%
Maintaining	*1.8%*	***34.6%***	*3.9%*	40.3%
Meets Goal	*0.0%*	*4.2%*	***49.8%***	54.0%
Total Assigned:	5.6%	40.7%	53.7%	100.0%
Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums. Total congruence (sum of bold numbers) = 88.1%				

Table 9.
Small High Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***8.7%***	*6.9%*	*0.2%*	15.9%
Maintaining	*2.9%*	***24.7%***	*8.4%*	36.0%
Meets Goal	*0.1%*	*7.6%*	***40.4%***	48.1%
Total Assigned:	11.8%	39.2%	49.0%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 73.7%

Table 10.
Medium High Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***12.1%***	*3.0%*	*0.0%*	15.1%
Maintaining	*2.2%*	***27.4%***	*5.7%*	35.2%
Meets Goal	*0.0%*	*7.3%*	***42.4%***	49.6%
Total Assigned:	14.3%	37.7%	48.1%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 81.9%

Table 11.
Large High Schools Expected Proportions of True Classifications Being in Each Possible Classification Given the Assigned Classifications

Possible True Classifications	Assigned Classification			Expected in True Classifications
	Needs Assistance	Maintaining	Meets Goal	
Needs Assistance	***5.2%***	*4.6%*	*0.0%*	9.8%
Maintaining	*1.5%*	***27.3%***	*5.4%*	34.2%
Meets Goal	*0.0%*	*3.6%*	***52.4%***	55.9%
Total Assigned:	6.7%	35.6%	57.8%	100.0%

Note: Bold numbers indicate proportions of all students with true classification and observed classification being congruent, by classification. Rounding may affect sums.
Total congruence (sum of bold numbers) = 84.9%

Classification accuracy by index

As mentioned above, classification errors are inevitable. In order to illustrate this concept, we have fine-tuned our analysis of errors by tracing probabilities that true classification is Assistance, Maintaining, or Meets Goal for accountability difference values from –7 to +7. Complete results are presented in the Technical Appendix, with the results for medium size elementary schools presented below. Based on the standard error of *prediction* from the regression analyses for the interim data, the dividing line between Assistance and Maintaining was set at –4.8. The dividing line between Maintaining and Meets Goal was set by regulation at 0. Figure 2 indicates that if a school were to have a calculated index of exactly –4.8 or exactly 0, there is essentially a fifty-fifty chance that they were correctly classified. On the other hand, for schools with calculated indexes that were one point away from their dividing line, the chances were approximately 80% that they were correctly classified. The further schools scores were from the nearest dividing line, the more likely they were to be correctly classified. We can also see in the figure that schools with observed index scores that qualified them for assistance (below –4.8) had essentially zero probability that their students' true achievements, perfectly measured, would have classified them as eligible for a reward.

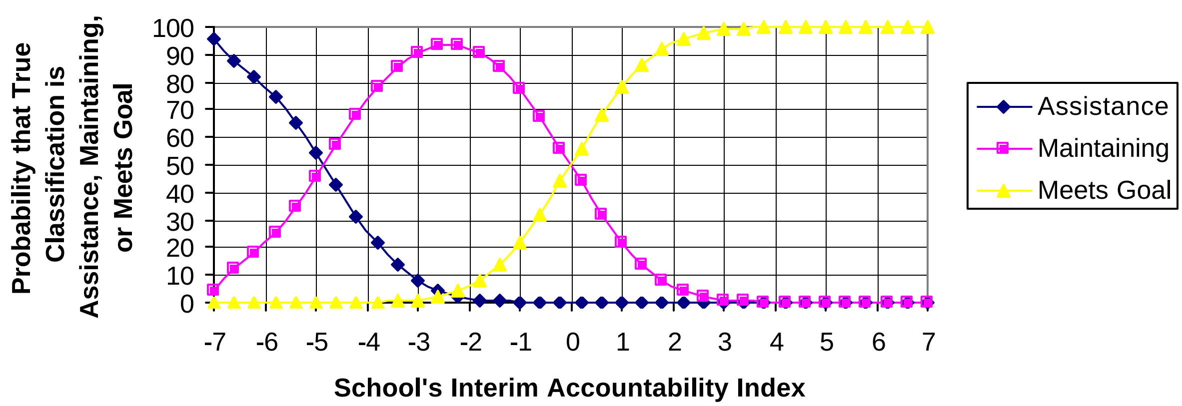

Figure 2. Medium Size Elementary Schools.

Perspective on Classification Accuracy

Given the nature of the data and the computations used to determine schools' accountability classifications, the results are about what might be expected. The following is excerpted from our student-level classification accuracy analysis (Hoffman & Wise, 2000c). It is applicable to the school-level analyses as well.

"Test specialists are currently in the early stages of recognizing the need to study classification accuracy as well as more traditional measures of test reliability. Currently, investigations of classification accuracy tend to be methodological papers which focus on analytical variations on the accuracy theme. It is instructive to examine several of these studies that use operational data. For example, Rogosa (1994) examined 1993 California's CLAS

assessment which uses six proficiency levels. He found that although the probability of classification within one category of true proficiency was nearly 95%, the probability of exact classification was only 51.72%. Rogosa (2000) has provided similar data for other assessments, including California's current assessment, STAR, along with a warning that test accuracy is often not as good as we think it is.

"In another example, Lee, Hanson, and Brennan (2000) used data from ACT's Work Keys assessment. Their results confirm that the number of proficiency categories makes a difference – more categories mean more opportunities for classification error and therefore less accuracy. For a Work Keys subtest with five categories, exact accuracy for several different forms was in the 70% range, while a subtest with six categories showed accuracy in the low- to mid-60% range. Lee et al also looked at accuracy for classifying students simply above or below a single cutpoint, and they used each of the possible Work Keys cutpoints to look at these dichotomous classifications. Accuracy was in the upper 80% range to near 100% for classifying students into only one of two categories. The higher levels of accuracy occurred for classification of students into either extreme. When the cutpoint was more near the center, accuracy tended to be in the upper 80% range. Young and Yoon (1998) provide some similar data from the New Standards assessments. Again, when making only a dichotomous (two category) classification, they showed accuracy in the lower 90% range."

For comparison purposes, we can calculate accuracy for the Interim Accountability model as if it had been used to divide schools into two categories – above improvement goal and below improvement goal. Looking at the data in Table 2 from this perspective, some of the cells that previously represented misclassification, now represent accurate classification. The resulting "dichotomous" accuracy of above versus below prediction is approximately 86%. Across all school levels and sizes, this dichotomous accuracy ranges from 83% to 93%, representing a 3% to 10% increase from the three-category results. Similar results were obtained when CATS student-level classifications were dichotomized by combining Novice with Apprentice and Proficient with Distinguished (Hoffman and Wise, 2000a and 2000b)

The school-level dichotomous results are in the same range as the Work Keys, New Standards Assessment, and the Kentucky Core Content Test. When comparing the CATS school accountability results to the classification accuracies of these individual student assessments, two differences are clear: 1) CATS accountability scores represent the aggregate performance of a significant number of students which would ordinarily be expected to increase accuracy over the student level classifications; 2) scores computed as differences are notoriously unreliable, a problem that is increased by using highly correlated data. Since the accountability classifications are in fact made on the basis of differences between scores from highly correlated data, CATS accountability scores must overcome a significant computational handicap. The combination of these two factors, which have opposite impact, results in the effects of the measurement error in the school classifications being comparable to the effects of measurement error in the student classifications.

Final Comment

As psychometricians, we have only presented the data. We will refrain from making policy statements about whether the accuracy in classification is acceptable given the benefits from the CATS initiative as a whole. We will note, however, that these data can be used to forecast the future accuracy of CATS classifications decisions. In the future, setting school targets will return to a simple school improvement concept, replacing the regression concept of comparisons between schools. However, future classification will still be based on differences between observed performance and projected performance. With a slight variation, computations used in the present analysis will be applicable to future CATS decisions and therefore, we can anticipate that future results will be in the same general range as the present results. A strength of the future decision system is that it will incorporate safety bands around performance targets that will be based on measurement analyses such as presented in this report.

References

Carlson, James E. (1999). *Kentucky accountability regressions 1997-98 to 1999 prediction. Monterey, CA: CTB/McGraw-Hill.*

Hoffman, R. G., & Wise, L. L. (1999). *Establishing the reliability of student level classifications: Analytic plan and demonstration.*(FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2000a). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications of the Kentucky Core Content Test* (FR-00-25). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2000b). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2000 Kentucky Core Content Test* (FR-00-41). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Wise, L. L. (2000c). *Establishing the reliability of student proficiency classifications: The accuracy of observed classifications.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000. Also available at www.Humrro.org.

Hoffman, R.G., & Wise, L.L. (2000d). *School classification accuracy final analysis plan for the commonwealth accountability and testing system* (FR-00-26). Alexandria, VA: Human Resources Research Organization.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000). *Procedures for computing classification consistency and accuracy indices with multiple categories.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000.

Miller, M. David. (April, 1999). *Generalizability of Performance-Based Assessments at the*

School Level. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1999.

Rogosa, D. (2000). *Statistical topics in educational assessment: Individual scores, Group summaries, and accountability systems*. Presented to the March 14, 2000 CCSSO Technical Issues in Large Scale Assessment Workshop, San Diego, California.

Rogosa, D. (1994). Misclassification in student performance levels. In CTB/McGraw-Hill. (1994). *1994 CLAS Assessment Technical Report*. Monterrey, CA: Author.

Yen, W. M. (1997). The technical quality of performance assessments: Standard Errors of Percents of Pupils Reaching Standards. *Educational Measurement: Issues and Practice, 16,* 5-15.

Young, M. J. & Yoon, B. (1998). *Estimating the Consistency and Accuracy of Classifications in a Standards-Referenced Assessment.* (CSE Technical Report 475). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Technical Appendix

The purpose of this appendix is to outline, for the technical reader, the procedures used to calculate classification accuracy. Selected intermediate results are also presented. There were nine broad steps to the process:

1. Identify target school sizes.
2. Select eligible schools and create school files by randomly selecting students within each school to meet targeted numbers.
3. Estimate standard errors of measurement using generalizability analyses for each grade/subject/school size combination for combined 1999 and 2000 data.
4. Calculate school-level correlations among accountability components for each school level.
5. Compute reliability and standard errors of measurement for school accountability index for each school level and size combination.
6. Adopt an estimate for standard error of measure for 1997 and 1998 school years for each school level and size combination.
7. Compute the standard error for the difference between the 1999/2000 index and the 1997/1998-based target index for each school level and size combination.
8. Using standard error estimates, compute accountability classification accuracy for each school level and size combination
9. Plot classification error curves.

Each of these steps is outline below.

1. Identify target school sizes

Because the number of students within a school will affect the reliability of school-level scores, three representative school sizes were included in the analysis. Because schools differ in the number of grades they contain, and because the analysis begins with grade-level data, we defined school size by the average number of students in a grade. Three sizes of schools were targeted at the elementary, middle, and high school levels. Small schools were identified as those in the smallest 1/3rd of all schools, and the representative size set at the median of that third, which is also the 16.7th percentile of all schools. Similarly, medium size schools were those in the middle 1/3rd and were represented by the 50th percentile of all schools. Finally, large schools were the largest 1/3rd and were represented by the 83.3th percentile of all schools. The selection was slightly complicated by needing to analyze data from different grades for two different years. That is, either the size of Grade 4 or Grade 5, for 1999 or 2000, or an average, could define elementary school percentiles. In fact, this wrinkle was superceded by a larger concern. The Kentucky Core Content Test is divided into multiple forms and we needed each of the different test forms to be represented equally in our analyses. Therefore, target sizes had to be adjusted to the nearest multiple of 12 – the number of Arts & Humanities and Practical Living/Vocational Studies forms. By using the 12 as the multiple, we also accommodated the 6 forms for the remaining subject areas. The table below shows the distribution of school sizes by grade and year. For reference, school sizes at the medians and the boundaries of the 1/3rd size divisions are indicated, along with the maximum size school.

Table A-1
Identification of Representative School Sizes

Grade	Year	School Sizes by Percentile					
		16.7th	33.3rd	50th	66.7th	88.3th	Maximun
4	1999	30	45	59	75	96	246
4	2000	29	47	61	76	96	255
5	1999	28	44	57	73	89	290
5	2000	30	46	59	75	94	291
	Elementary target	**24**		**60**		**96**	
7	1999	35	70	126	191	246	438
7	2000	36	67	127	190	259	459
8	1999	36	71	133	191	256	430
8	2000	36	70	126	194	247	423
	Middle target	**36**		**120**		**240**	
10	1999	61	115	179	228	298	624
10	2000	63	119	173	222	292	644
11	1999	65	110	164	202	258	563
11	2000	65	110	163	206	261	518
	High School target	**60**		**168**		**240**	

2. Select eligible schools and create school files by randomly selecting students within each school to meet targeted numbers.

Given that there are not schools with exactly these target numbers of students and with an exactly equal representation of subject forms, the next step was to create synthetic schools with exactly the target representation. This was done by randomly selecting/eliminating students from existing schools. However before this random selection of students could begin, candidate schools had to be identified. Because small, medium, and large size schools have characteristics other than size that may affect measurement accuracy (e.g., smaller schools may be more homogeneous), only schools near the target size were considered eligible for the analyses. Certainly, schools could be no smaller than the target size. Selection of the maximum size became a trial and error process. We discovered that the criteria for having equal numbers of forms led to the need to consider larger schools for the maximum size than we originally expected. Table A-2 indicates the ranges of school sizes, from target size to maximum size, that became candidates for our analyses and the numbers of such schools. In each case, we tried to balance having enough schools for stable generalizability results without having the maximum size being subjectively larger than the target size. This was most difficult to achieve for the small size middle and high schools.

Table A-2.
Ranges of candidate school sizes and numbers of schools in those ranges

Level	Small			Medium			Large		
	Target Size	Max. Size	No. of Schools	Target Size	Max. Size	No. of Schools	Target Size	Max. Size	No. of Schools
Elementary									
Grade 4	24	36	53	60	78	80	96	120	52
Grade 5	24	36	50	60	78	81	96	120	42
Middle									
Grade 7	36	60	34	120	170	26	240	360	47
Grade 8	36	60	29	120	170	31	240	360	51
High School									
Grade 10	60	120	33	168	240	43	240	644	69
Grade 11	60	120	44	168	240	41	240	644	48
Grade 12	60	120	42	168	240	49	240	644	36

Having identified eligible schools from which to create schools of the exact target sizes, the next step was a straightforward random selection of appropriate numbers of students. The requirement of having equal numbers of forms for the analysis frequently eliminated schools near the target because there were too few students for one or more forms. The generalizability results that follow will show the numbers of schools actually used.

3. Estimate standard errors of measurement using generalizability analyses for each grade/subject/school size combination for combined 1999 and 2000 data.

Figure A-1 presents the design and Appendix Tables A-3, A-4 and A-5 present the formulas for all subjects (except Writing Portfolios) using Brennan's (1981) notation for generating sums of squares and variance components. For each of the grade/subject combinations, six sources of variance in schools' two-year academic index averages include: (1) school, (2) year, (3) school by year, (4) form within year, (5) school by form within year, and (6) pupil within form within school by year. The order of the nesting terms in the last source of variance is a little ambiguous in its wording since pupils are nested within forms, within schools, and within years. However, for derivation of the error components, the expressed order of the nesting does not matter, as long as the nesting is captured. For the Writing Portfolio, there is no form component and Brennan presents formulas that include this case. Results are presented in Table A-6.[4]

[4] Note that in Brennan's presentation, "persons" are the objects of measurement and therefore variables with the subscript "p" refer to the objects of measurement. In our case, schools, noted by the subscript "s," are the objects of measurement. In our notation, variables with the subscript "p" refer to pupils as one facet of the school scores.

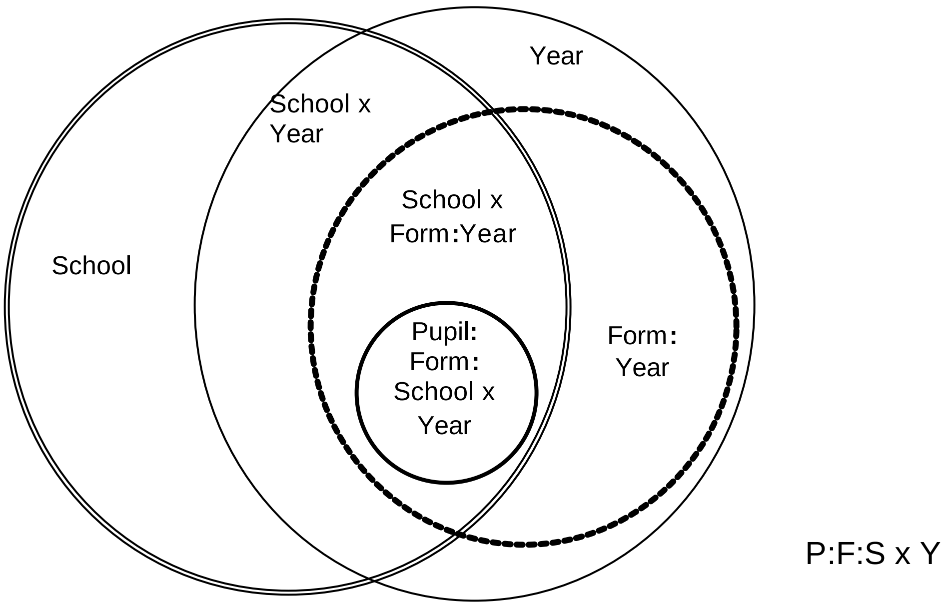

Figure A-1. Generalizability theory design representing Kentucky Core Content Test two-year accountability cycle.

Table A-3.
Random Effects Variance Components Estimates for Pupil: School Year Form Generalizability Theory Design

Effect	df	Means	SS
School (s)	$n_s - 1$	$\bar{X}_s = \frac{1}{n_y n_f n_p} \sum_y \sum_f \sum_p X_{syfp}$	$n_f n_y n_p \Sigma \bar{X}_s^2 - n_s n_y n_f n_p \bar{X}^2$
Year (y)	$n_y - 1$	$\bar{X}_y = \frac{1}{n_s n_f n_p} \sum_s \sum_f \sum_p X_{syfp}$	$n_s n_f n_p \Sigma \bar{X}_y^2 - n_s n_y n_f n_p \bar{X}^2$
School x Year	$(n_s - 1)(n_y - 1)$	$\bar{X}_{sy} = \frac{1}{n_f n_p} \sum_f \sum_p X_{syfp}$	$n_f n_p \Sigma\Sigma \bar{X}_{sy}^2 - n_f n_y n_p \Sigma \bar{X}_s^2 - n_s n_f n_p \Sigma \bar{X}_y^2 + n_s n_y n_f n_p \bar{X}^2$
Form: Year (f:y)	$n_y(n_f - 1)$	$\bar{X}_{f:y} = \frac{1}{n_s n_p} \sum_s \sum_p X_{syfp}$	$n_s n_p \Sigma\Sigma \bar{X}_{yf}^2 - n_s n_f n_p \Sigma \bar{X}_y^2$
School x Form : Year (sf:y)	$n_y(n_s - 1)(n_f - 1)$	$\bar{X}_{sf:y} = \frac{1}{n_p} \sum_p X_{syfp}$	$n_p \Sigma\Sigma\Sigma \bar{X}_{syf}^2 - n_f n_p \Sigma\Sigma \bar{X}_{sy}^2 - n_s n_p \Sigma\Sigma \bar{X}_{yf}^2 + n_s n_f n_p \bar{X}_y^2$
Pupil: School Year Form (p:sfy)	$n_y n_s n_f (n_p - 1)$	na	$\Sigma\Sigma\Sigma\Sigma X_{psyf}^2 - n_p \Sigma\Sigma\Sigma \bar{X}_{syf}^2$
Total	$n_s n_y n_f n_p - 1$	$\bar{X} = \frac{1}{n_s n_y n_f n_p} \sum_s \sum_y \sum_f \sum_p X_{syfp}$	$\Sigma\Sigma\Sigma\Sigma X_{psyf}^2 - n_s n_y n_f n_p \bar{X}^2$

Table A-4.
G-Study Variance Components Estimates for Pupil: School Year Form Generalizability Theory Design

Effect (α)	Estimated σ^2 –Random Effects Model	Estimated $\sigma^2(\alpha\|M)$ -- Mixed Models (N = Universe size)	
		Basic Mixed Model	Year Fixed
School (s)	$\frac{[MS(s) - MS(sy)]}{n_y n_f n_p}$	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y} + \frac{\hat{\sigma}^2_{sf:y}}{N_f N_y} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_y N_p}$	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$
Year (y)	$\frac{[MS(y) - MS(sy) - MS(fy) + MS(sfy)]}{n_s n_f n_p}$	$\hat{\sigma}^2_y + \frac{\hat{\sigma}^2_{sy}}{N_s} + \frac{\hat{\sigma}^2_{f:y}}{N_f} + \frac{\hat{\sigma}^2_{sf:y}}{N_s N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_f N_p}$	$\hat{\sigma}^2_y$
School x Year	$\frac{[MS(sy) - MS(sfy)]}{n_f n_p}$	$\hat{\sigma}^2_{sy} + \frac{\hat{\sigma}^2_{sf:y}}{N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_p}$	$\hat{\sigma}^2_{sy}$
Form: Year (f:y)	$\frac{[MS(fy) - MS(sfy)]}{n_s n_p}$	$\hat{\sigma}^2_{f:y} + \frac{\hat{\sigma}^2_{sf:y}}{N_s} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_p}$	$\hat{\sigma}^2_{f:y}$
School x Form : Year (sf:y)	$\frac{[MS(sfy) - MS(syfp)]}{n_p}$	$\hat{\sigma}^2_{f:sy} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_p}$	$\hat{\sigma}^2_{f:sy}$
Pupil: School Year Form (p:sfy)	MS(syfp)	$\hat{\sigma}^2_{p:f:sy}$	$\hat{\sigma}^2_{p:f:sy}$

Table A-5.
D-Study Variance Components Estimates for Pupil: School Year Form Generalizability Theory Design

Effect (α)	D-study error component	Use term in	
		Absolute error estimate	Relative error estimate
School (s)	$\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$		
Year (y)	$[\hat{\sigma}^2_y / N_y]\ [1 - \frac{n_y}{N_y}] = \mathbf{0}$	(X)	
School x Year	$[\hat{\sigma}^2_{sy} / N_y] \times [1 - \frac{n_y}{N_y}] = \mathbf{0}$	(X)	(X)
Form: Year (f:y)	$\hat{\sigma}^2_{f:y} / N_y N_f$	X	
School x Form : Year (sf:y)	$\hat{\sigma}^2_{f:sy} / N_y N_f$	X	X
Pupil: School Year Form (p:sfy)	$\hat{\sigma}^2_{p:f:sy} / N_y N_f n_p$	X	X

Table A-6.
Generalizability Theory Results

rd = Reading sc = Science wo = Writing Prompt wp = Writing Portfolio ah = Arts & Humanities ma = Mathematics pl = PL/VS ss = Social Studies		Lg = Large Md = Medium Sm = Small	NS = Number of Schools NP = Number of Pupils NF = Number of Forms NY = Number of Years				Ab, Err = Absolute Error Variance Rel. Error = Relative Error Variance Tot Var. = Total Variance Ab. Gen. = Absolute Generalizability Rel. Gen. = Relative Generalizability			
Grade	Subject	Size	NS	NP	NF	NY	Ab. Error	Rel. Err	Tot Var.	Ab. Gen.	Rel. Gen.
4	rd	lg	36	16	6	2	3.02	2.85	52.33	0.94	0.95
4	rd	md	55	10	6	2	4.03	4.03	67.39	0.94	0.94
4	rd	sm	44	4	6	2	9.90	9.51	37.36	0.74	0.75
4	sc	lg	36	16	6	2	2.86	2.83	50.85	0.94	0.94
4	sc	md	55	10	6	2	3.57	3.57	85.02	0.96	0.96
4	sc	sm	44	4	6	2	7.99	7.94	46.49	0.83	0.83
4	wo	lg	35	16	6	2	5.65	5.51	44.07	0.87	0.87
4	wo	md	54	10	6	2	7.97	7.90	52.79	0.85	0.85
4	wo	sm	42	4	6	2	15.89	15.87	47.13	0.66	0.66
4	wp	lg	54	96	-	2	4.05	4.05	147.10	0.97	0.97
4	wp	md	29	60	-	2	6.09	6.09	199.89	0.97	0.97
4	wp	sm	51	24	-	2	17.68	17.68	227.60	0.92	0.92
5	ah	lg	28	8	12	2	6.19	6.05	95.85	0.94	0.94
5	ah	md	39	5	12	2	8.18	7.93	75.50	0.89	0.89
5	ah	sm	39	4	12	2	6.67	6.51	42.03	0.84	0.85
5	ma	lg	33	16	6	2	7.86	7.59	230.57	0.97	0.97
5	ma	md	57	10	6	2	10.23	10.23	207.20	0.95	0.95
5	ma	sm	39	4	6	2	23.85	23.85	186.05	0.87	0.87
5	pl	lg	28	8	12	2	4.78	4.65	66.08	0.93	0.93
5	pl	md	38	5	12	2	6.84	6.79	57.02	0.88	0.88
5	pl	sm	28	2	12	2	15.06	15.06	65.03	0.77	0.77
5	ss	lg	32	16	6	2	3.92	3.92	108.09	0.96	0.96
5	ss	md	57	10	6	2	5.87	5.72	98.18	0.94	0.94
5	ss	sm	39	4	6	2	13.34	13.02	84.06	0.84	0.85
7	rd	lg	41	40	6	2	0.78	0.76	47.14	0.98	0.98
7	rd	md	22	20	6	2	1.68	1.65	16.70	0.90	0.90
7	rd	sm	28	6	6	2	4.66	4.66	46.59	0.90	0.90
7	sc	lg	41	40	6	2	0.80	0.80	37.47	0.98	0.98
7	sc	md	22	20	6	2	1.40	1.37	16.66	0.92	0.92
7	sc	sm	28	6	6	2	3.19	3.15	38.59	0.92	0.92
7	wo	lg	41	40	6	2	2.51	2.26	64.10	0.96	0.96
7	wo	md	22	20	6	2	4.33	4.25	30.43	0.86	0.86
7	wo	sm	27	6	6	2	11.79	11.79	75.78	0.84	0.84
7	wp	lg	48	240	-	2	1.73	1.73	148.41	0.99	0.99
7	wp	md	27	120	-	2	3.70	3.70	69.19	0.95	0.95
7	wp	sm	36	36	-	2	12.67	12.67	120.42	0.89	0.89
8	ah	lg	29	20	12	2	2.39	2.28	77.73	0.97	0.97
8	ah	md	27	10	12	2	4.52	4.33	74.54	0.94	0.94

Table A-6.
Generalizability Theory Results

rd = Reading sc = Science wo = Writing Prompt wp = Writing Portfolio ah = Arts & Humanities ma = Mathematics pl = PL/VS ss = Social Studies	Lg = Large Md = Medium Sm = Small	NS = Number of Schools NP = Number of Pupils NF = Number of Forms NY = Number of Years	Ab, Err = Absolute Error Variance Rel. Error = Relative Error Variance Tot Var. = Total Variance Ab. Gen. = Absolute Generalizability Rel. Gen. = Relative Generalizability

Grade	Subject	Size	NS	NP	NF	NY	Ab. Error	Rel. Err	Tot Var.	Ab. Gen.	Rel. Gen.
8	ah	sm	21	3	12	2	12.03	11.91	197.14	0.94	0.94
8	ma	lg	40	40	6	2	3.44	3.44	186.43	0.98	0.98
8	ma	md	28	20	6	2	6.87	6.75	105.76	0.94	0.94
8	ma	sm	26	6	6	2	19.23	19.23	462.74	0.96	0.96
8	pl	lg	30	20	12	2	2.30	2.25	44.95	0.95	0.95
8	pl	md	27	10	12	2	4.71	4.71	52.97	0.91	0.91
8	pl	sm	20	3	12	2	12.84	12.84	174.49	0.93	0.93
8	ss	lg	41	40	6	2	2.17	2.16	74.66	0.97	0.97
8	ss	md	28	20	6	2	3.42	3.28	72.79	0.95	0.95
8	ss	sm	26	6	6	2	8.83	8.72	193.60	0.95	0.95
10	pl	lg	47	20	12	2	2.34	2.26	66.73	0.96	0.97
10	pl	md	29	14	12	2	3.43	3.34	37.79	0.91	0.91
10	pl	sm	26	5	12	2	8.00	8.00	35.43	0.77	0.77
10	rd	lg	56	40	6	2	2.07	2.00	72.55	0.97	0.97
10	rd	md	39	28	6	2	3.43	3.31	50.79	0.93	0.93
10	rd	sm	29	10	6	2	6.53	6.53	55.25	0.88	0.88
11	ah	lg	35	20	12	2	2.22	2.08	91.42	0.98	0.98
11	ah	md	24	14	12	2	2.63	2.51	60.40	0.96	0.96
11	ah	sm	34	5	12	2	6.44	6.44	57.63	0.89	0.89
11	ma	lg	40	40	6	2	3.26	3.08	168.95	0.98	0.98
11	ma	md	27	28	6	2	4.26	4.17	185.55	0.98	0.98
11	ma	sm	38	10	6	2	10.93	10.88	122.55	0.91	0.91
11	sc	lg	40	40	6	2	1.16	1.07	42.25	0.97	0.97
11	sc	md	27	28	6	2	1.57	1.42	36.79	0.96	0.96
11	sc	sm	38	10	6	2	3.92	3.86	33.57	0.88	0.88
11	ss	lg	40	40	6	2	2.36	2.30	116.18	0.98	0.98
11	ss	md	27	28	6	2	3.01	2.95	91.49	0.97	0.97
11	ss	sm	38	10	6	2	8.10	8.10	72.99	0.89	0.89
12	wo	lg	29	40	6	2	1.67	1.61	21.86	0.92	0.93
12	wo	md	29	28	6	2	2.85	2.64	37.94	0.92	0.93
12	wo	sm	29	10	6	2	6.26	6.26	40.97	0.85	0.85
12	wp	lg	36	240	-	2	1.99	1.99	61.67	0.97	0.97
12	wp	md	50	168	-	2	3.00	3.00	82.67	0.96	0.96
12	wp	sm	42	60	-	2	7.96	7.96	92.52	0.91	0.91

4. Calculate school-level correlations among accountability components for each school level, elementary, middle, and high school.

Correlations were calculated separately for each school level, but not by school size. Tables A-7, A-8, and A-9 present the correlations.

Table A-7.
Correlations among accountability components for High Schools

	Reading	Math	Science	Soc Stud	A&H	PL/VS	Prompt	Portfolio	N-A
Mean	67.59	65.94	66.12	66.50	38.81	37.93	47.05	57.05	94.48
SD	9.13	12.50	6.84	10.22	8.11	7.18	6.68	9.46	2.55
N	239	239	239	239	239	239	239	239	237
Correlations									
Mathematics	0.834								
Science	0.861	0.902							
Social Studies	0.882	0.898	0.925						
Art & Humanities	0.828	0.859	0.853	0.925					
PL/VS	0.893	0.842	0.820	0.852	0.843				
Writing Prompt	0.661	0.688	0.676	0.681	0.684	0.628			
Writing Portfolio	0.554	0.605	0.561	0.573	0.542	0.504	0.533		
Non-Academic	0.587	0.656	0.603	0.613	0.536	0.568	0.496	0.416	

Table A-8.
Correlations among accountability components for Middle Schools

	Reading	Math	Science	Soc Stud	A&H	PL/VS	Prompt	Portfolio	N-A
Mean	67.03	67.06	36.62	51.14	39.92	32.22	40.33	40.28	96.60
SD	6.28	14.54	5.27	8.90	9.52	7.57	7.70	10.84	1.48
N	342.00	338.00	342.00	338.00	338.00	338.00	342.00	342.00	341.00
Correlations									
Mathematics	0.849								
Science	0.895	0.853							
Social Studies	0.861	0.903	0.862						
Art & Humanities	0.833	0.856	0.828	0.916					
PL/VS	0.807	0.856	0.829	0.906	0.896				
Writing Prompt	0.836	0.765	0.805	0.810	0.787	0.753			
Writing Portfolio	0.591	0.515	0.566	0.581	0.570	0.546	0.612		
Non-Academic	0.463	0.423	0.466	0.419	0.384	0.391	0.399	0.339	

Table A-9.
Correlations among accountability components for Elementary Schools

	Reading	Math	Science	Soc Stud	A&H	PL/VS	Prompt	Portfolio	N-A
Mean	76.04	63.92	55.56	56.64	31.21	44.04	33.60	57.17	95.78
SD	7.88	13.76	8.07	9.48	8.96	8.13	7.56	13.28	0.99
N	782.00	772.00	782.00	772.00	772.00	772.00	782.00	782.00	794.00
Correlations									
Mathematics	0.776								
Science	0.916	0.738							
Social Studies	0.806	0.887	0.773						
Art & Humanities	0.711	0.801	0.687	0.851					
PL/VS	0.766	0.843	0.732	0.897	0.838				
Writing Prompt	0.792	0.646	0.752	0.656	0.612	0.651			
Writing Portfolio	0.549	0.427	0.549	0.471	0.441	0.461	0.590		
Non-Academic	0.569	0.577	0.529	0.531	0.484	0.547	0.457	0.358	

5. Compute reliability and standard errors of measurement for school accountability index for each school level and size combination.

The school accountability index is a weighted composite of the eight Kentucky Core Content Test components and the Non-Academic Index, where each of the eight components include two years of data. Given the above estimates of total and error variance (using absolute error), the weights (set by regulation), and the correlations among the components, standard formulas for calculating the variance of a composite were applied to produce estimates for total variance and error variance of the two-year school accountability composite. This composite is call "Component 1," according to state regulations. The resulting estimates are presented in Table A-10.

Table A-10.
Component 1 Error Variance and Reliability

School Level and Size	Total Variance	Error Variance	Reliability
Large High Schools	58.534	0.275	.995
Medium High Schools	51.728	0.375	.993
Small High Schools	44.877	0.884	.980
Large Middle Schools	53.848	0.222	.996
Medium Middle Schools	33.833	0.416	.988
Small Middle Schools	98.884	1.148	.988
Large Elementary Schools	71.160	0.679	.990
Medium Elementary Schools	77.068	0.916	.988
Small Elementary Schools	61.332	2.154	.965

6. Adopt an estimate for standard error of measurement for 1997 and 1998 school years for each school level and size combination.

When KIRIS transitioned into CATS between 1998 and 1999, a number of test format differences were introduced. The differences precluded use of the generalizability approach described above on the 1997 and 1998 test data. As a result, we did not have a method for directly estimating the error variance in the 1997 and 1998 data used to establish school goals. Therefore, we adopted the following strategy. With the addition of the multiple choice items, the 1999 and 2000 tests included twice as many score points as the 1997 and 1998 tests. Therefore, to estimate the reliability of the 1997/1998 school composite, we applied the Spearman-Brown prophecy formula to the 1999/2000 composite reliability to produce a reliability estimate for a test that is one-half as long.

7. Compute the standard error for the difference between the 1999/2000 index and the 1997/1998-based target index for each school level and size combination.

The interim accountability model classified schools according whether their 1999/2000 accountability index scores met target scores based on their 1997/1998 accountability scores. In effect, the classification was based on the residual of schools' actual 1999/2000 index compared to the index predicted by their 1997/1998 index, given the linear relationship between 1999/2000 scores and 1997/1998 scores. Since the residual is a weighted composite, the reliability of the composite could be estimated using the above variance and reliability estimates and the slope of the linear regression between 1998/1998 and 1999/2000 scores. Table A-11 shows the results.

Table A-10.

School Level and Size	Reliability estimate for 1997/1998 index	Regression slope between 1997/1998 and 1999/2000 indexes	Total Variance of Residual	Error Variance of Residual	Reliability of Residual
Large High Schools	.990	.894	9.88	0.65	.934
Medium High Schools	.986	.894	9.88	1.00	.899
Small High Schools	.961	.894	9.88	2.68	.727
Large Middle Schools	.992	.960	7.84	0.52	.932
Medium Middle Schools	.976	.960	7.84	1.57	.800
Small Middle Schools	.977	.960	7.84	1.48	.811
Large Elementary Schools	.981	.965	23.00	1.36	.941
Medium Elementary Schools	.976	.965	23.00	1.69	.926
Small Elementary Schools	.932	.965	23.00	4.93	.786

8. Using standard error estimates compute accountability classification accuracy for each school level and size combination.

The method used for estimating classification accuracy is described in detail in Hoffman and Wise (1999) and Hoffman and Wise (2000c). Briefly, the logic of the method is to use standard errors of measurement to compute a matrix of probabilities of various levels of observed scores given possible true scores, and then apply Bayes' theorem, coupled with an estimate of the true score distribution, to transform that initial matrix into a matrix of probability of alternative true scores given various possible observed scores.

9. Plot classification error curves.

Using the data from the matrix of probability of alternative true scores given various possible observed scores, classification accuracy was plotted as a function of observed scores. Figures A-2 through A-10 present these plots.

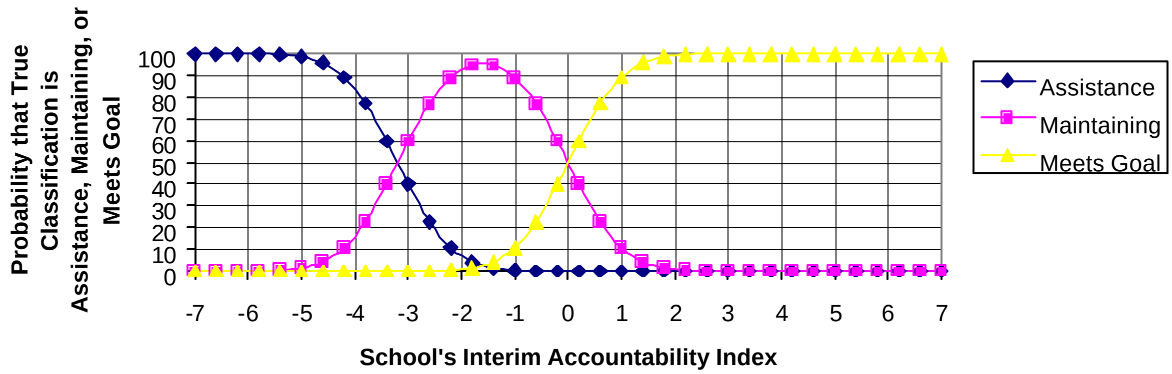

Figure A-2. Large High Schools.

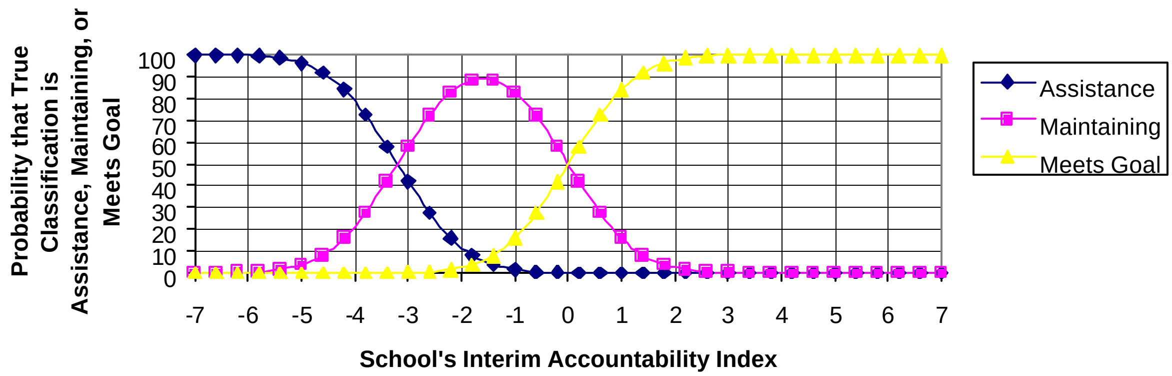

Figure A-3. Medium Size High Schools.

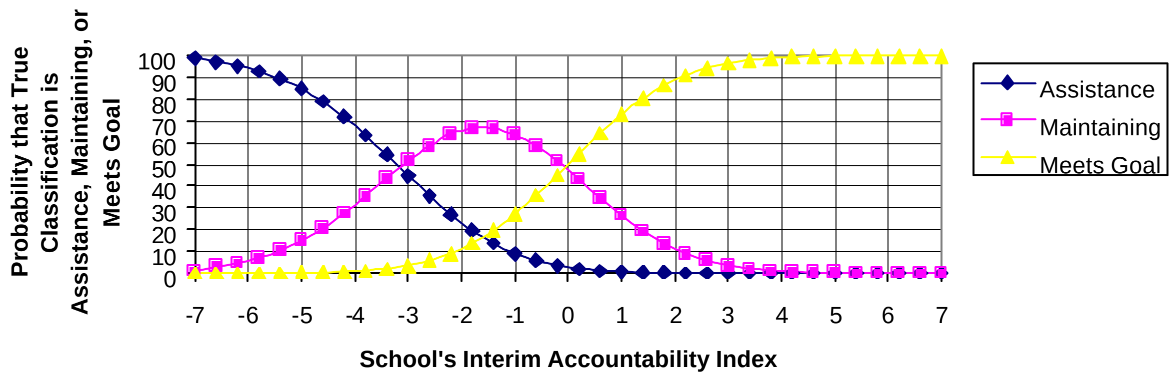

Figure A-4. Small HIgh Schools.

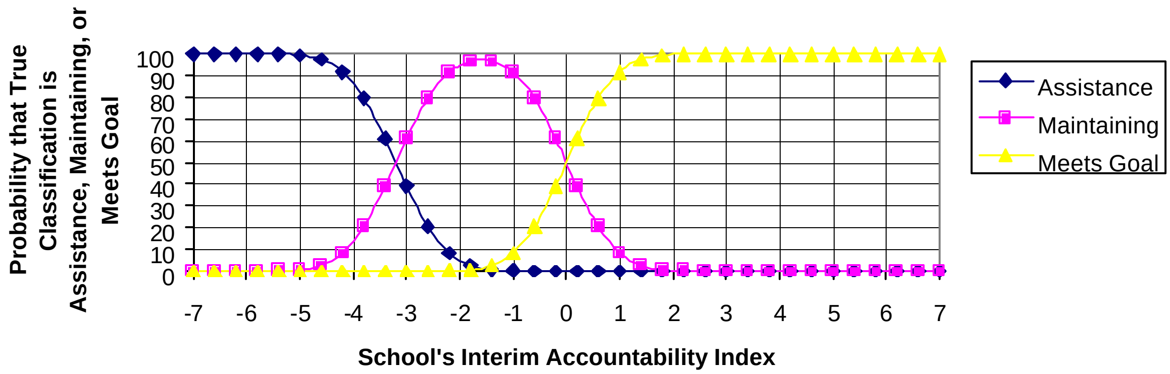

Figure A-5. Large Middle Schools.

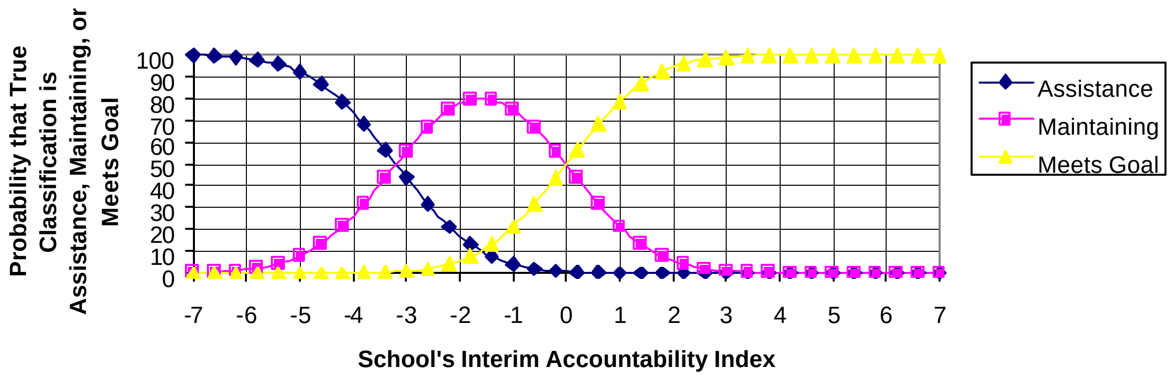

Figure A-6. Medium Size Middle Schools.

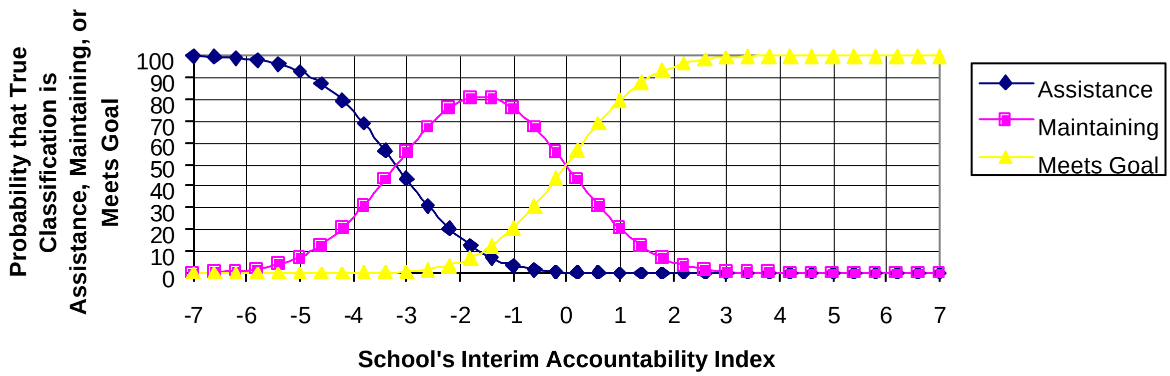

Figure A-7. Small Middle Schools.

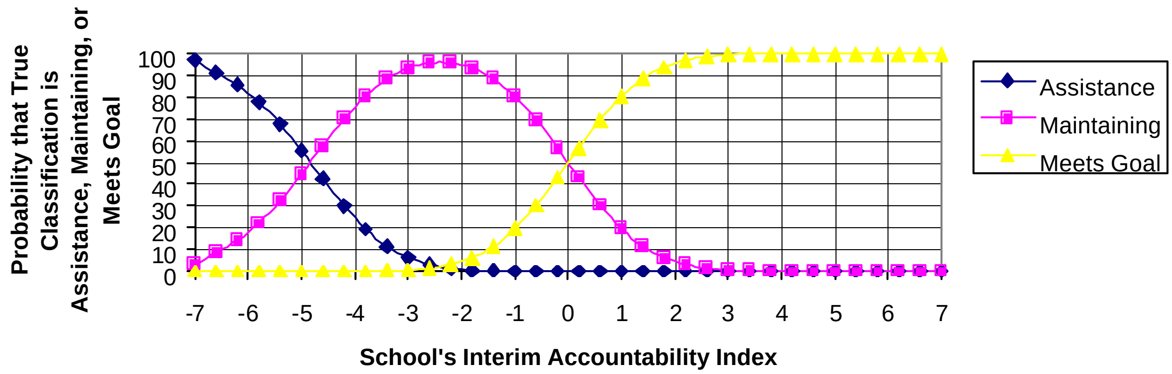

Figure A-8. Large Elementary Schools.

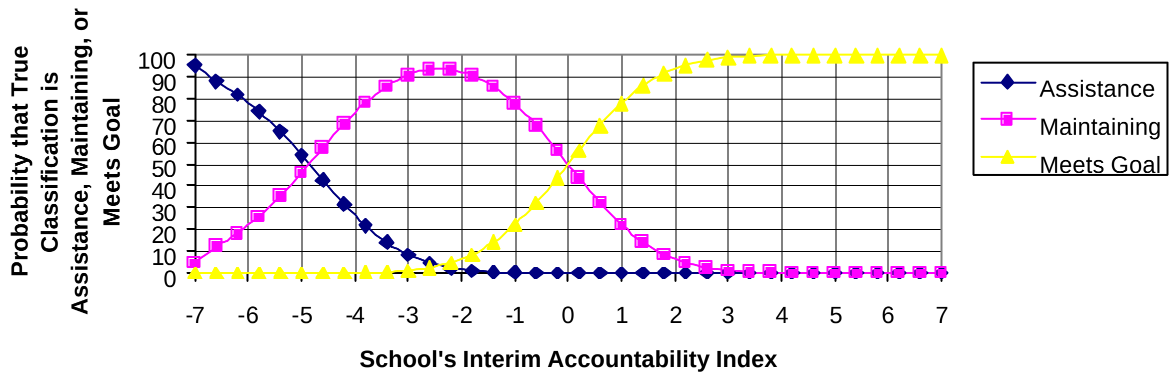

Figure A-9. Medium Size Elementary Schools.

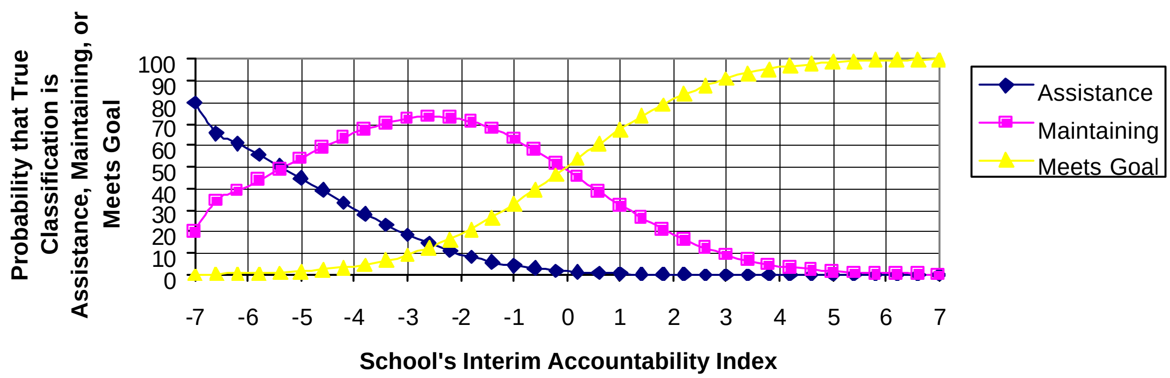

Figure A-10. Small Elementary Schools.